# Supporting Online Material

## Materials and Methods

*1. Generation of sequencing targets*. Sequencing targets were amplified from genomic DNA using Platinum Taq HiFi (Invitrogen) following the manufacturer's recommendations, with primers described below. Single-band PCR products were gel-purified (Gel-extraction kit, Qiagen) and subjected to direct microsequencing using custom primers. Fluorescence automated DNA sequencing was carried out using BygDye chemistry in an ABI3700 Sequencer (Applied Biosystems). Base calling, quality assessment and assembly was carried out using the phred, Phrap, phrapview, Consed software suite developed by Phil Green (www.phrap.org). Both the (+) and (-) strands were sequenced at least once. All the sequences generated in this study have been submitted to GenBank (acc. #: AY190030-AY190042, AY192729-AY192785).

*2. Phylogenetic analysis of the sequence data.* The various primate sequences for each gene were aligned using CLUSTALW (S1), and maximum likelihood phylogenetic trees were generated from the resulting alignments using fastDNAml (S2). Mutation rates under the Hasegawa-Kishino-Yano (HKY) model (S3) for "conserved" and "non-conserved" regions were learned by training on a separate data set. Gaps were treated as mutations by replacing each gap in a given column with the least common base in that column and breaking ties according to global base frequencies. The training set for the mutation rates comprised sequences from a set of 18 primate species containing the apolipoprotein(a) exon 1, using the known location of that exon. These sequences were generated to identify the primate-specific regulatory elements of this gene. Assigning a default rate of 1 for the coding (defined as "conserved") region, we found the mutation rate for the whole non-coding region to be 7.3, indicating that, as expected, non-coding regions evolve on average faster than coding regions. These rates were employed to calculate the likelihood ratio under a fast- versus a slow-mutation regime for each column of the multiple alignment across all four regions analyzed.

*3. Analysis of the discriminative power of phylogenetic shadowing at increasing species subset sizes.* The discriminative power of any fixed sub-collection of a set of species is directly related to the overall number of accumulated mutations in that collection's family tree (a sub-tree of the

phylogenetic tree for all species under consideration). The expected number of accumulated mutations, in turn, is determined by the family tree's weight, that is, the sum of its edge lengths. For each species subset size k, the subtree maximizing the tree weight among all minimal subtrees containing k leaves is called the k-MST. Thus, the 2-MST is the path of maximal weight from one leaf to another in the tree. It can be shown that the k-MSTs are efficiently computable using a recursive dynamic programming algorithm.

*4. Electrophoretic mobility shift patterns of the conserved and non-conserved intervals of the apo(a) 1.6 kb region.* Oligonucleotide probes (sequences are given below) were end-labeled with 32P-dCTP (Amersham). HepG2 nuclear protein extracts were isolated by lising cells in lysis buffer (0.1% NP-40, 1 mM DTT, 10 mM Hepes, pH 7.8, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA supplemented with protease inhibitors) and resuspending in 1 mM DTT, 20 mM Hepes, pH 7.8, 0.4 M NaCl, 1 mM EDTA, 1 mM EGTA. The binding reaction contained 1-4 g nuclear protein extract, 0.2-0.5 ng radiolabeled probe, 10 mM Hepes, pH 7.8, 1 mM EDTA, 1 mM EGTA, 12% glycerol, 0.06-0.12 mg/ml dI-dC polymer. Following a 15 min. incubation at room temperature, the reaction was analyzed by electrophoresis in a non-denaturing 5% acrylamide gel at 4°C in 0.5x TBE buffer. Autoradiograms were scanned with a Flour-S-MultiImager and analyzed the MultiAnalyst software (Bio Rad).

*5. Generation of apo(a)'s 1.6 bk region reporter constructs.* Apo(a)'s 1.6 kb region was amplified from the RP1-81d8 clone (available from www.chori.org) with the primers apo(a).16.Sma and apo(a).16.Nco (see below) using Platinum Taq HiFi (Invitrogen) following the manufacturer's recommendations. The PCR product was gel-purified (Gel-extraction kit, Qiagen), digested with SmaI and NcoI and ligated into a SmaI- and NcoI-cut pGL3-Basic vector (Promega) to generate the whole apo(a) luciferase reporter construct (pGL3.RP1). Deletion constructs were generated following the mismatch-primer protocol (S4), using pGL3.RP1 as template with the primers described in the Supporting online material. All deletion constructs were verified by DNA sequencing.

*6. Transfection analysis of apo(a)'s reporter constructs.* The human HepG2 hepatoma cells (ATCC HB8065) (the liver is the sole site of apo(a)'s expression) were grown in MEM (GibcoBRL) supplemented with 10%FBS (Hyclone), 2 mM glutamine, 0.1 mM minimal essential medium non-

essential amino acids, 1 mM sodium pyruvate. Approximately 2 × 105 cells/well were seeded in a 12-well cell culture plate 24 h prior to transfection by the Fugene method (Roche Molecular Biochemicals) following the manufacturer's protocol. Transfections were carried out in quadruplicates. Briefly, 2.0 μg of test-plasmid and 0.5 μg of pSV- β-galactosidase (Promega) control plasmid were mixed with 6 μl Fugene and added to each well and covered with 2 ml of maintenance medium. Following a 48-h incubation, the cells were harvested and lysed. Luciferase and β-galactosidase expression were assayed with the Bright-Glo Luciferase Assay System (Promega) and the β-Galactosidase Enzyme Assay System (Promega), respectively.

**Primer sequences**

1. Forward (f) and reverse (r) primers for the amplification of the apoB exon-19, CETP exon-8, PLG exon-6 and apo(a) exon-1 genomic intervals. Primers apo(a).2 are internal to primers apo(a).1.

| | |
|---|---|
| fApoB19 | CATTGATGAAAGCRTTTTCAGR |
| rApoB19 | AACATTGTCTCTTGCCCTTAACT |
| fCETP8 | GAACTCAGGACARAYGGGTGA |
| rCETP8 | CTGTGAGCAGAGTGGGTGTG |
| fLXRA3 | TCTTCAGGCGGATCTGTTCT |
| rLXRA3 | CTCAGAGAGGAAGCCAGGAT |
| fPLG6 | GAAGCGCTGTCACTTCAGGT |
| rPLG6 | CCCCTTAGAGAATCTGATGGAA |
| f.apo(a).1 | ACTTACATTCACAATCACACAC |
| r.apo(a).1 | TGCCATAACTACCTCAGACC |
| f.apo(a).2 | GGCAAATGTACACCAATGGAAAG |
| r.apo(a).2 | GGAGACTGGAGCTCAGATATTGC |

2. Oligonucleotides used to generate the duplexes analyzed by electrophoretic mobility shift assay. Primer X were annealed to primer Xrc (e.g. primer C1 was annealed to primer C1rc) prior to the labeling reaction.

| | |
|---|---|
| C1 | AGCTGGATTACAGGTGCCCACCACCACGCCTGGCTAATTTTTGTATTTTTAGTAGAGATGGGGTT |
| C1rc | AGCTAACCCCATCTCTACTAAAAATACAAAAATTAGCCAGGCGTGGTGGTGGGCACCTGTAATCC |
| C2 | AGCTACCTGTCTTGGCCTCCCAAAGTGCTGGGATTACAGAGTTG |
| C2rc | AGCTCAACTCTGTAATCCCAGCACTTTGGGAGGCCAAGACAGGT |
| C3 | AGCTGGAGTGCAGTGGCACATTCTTGGCTCACTGCAACCT |
| C3rc | AGCTAGGTTGCAGTGAGCCAAGAATGTGCCACTGCACTCC |
| C4 | AGCTTGCCTCCACACCCAGCTAATTTTTGTATTTTTTGTAGAGACAGG |
| C4rc | AGCTCCTGTCTCTACAAAAAATACAAAAATTAGCTGGGTGTGGAGGCA |
| C5 | AGCTCATGTTGCCCAGGCTTGTCTCAAACTCCTGGCCTCAGGTGATCCATCTGCCGTGG |
| C5rc | AGCTCCACGGCAGATGGATCACCTGAGGCCAGGAGTTTGAGACAAGCCTGGGCAACATG |
| C6 | AGCTACATAATTTATAATTACAGAAAACATGTGAGTTCACTAG |
| C6rc | AGCTCTAGTGAACTCACATGTTTTCTGTAATTATAAATTATGT |
| C7 | AGCTTATGGATGTGCAGCACTTGGCAGAGGGTCTGGTCATGGAAGTTACC |

| | |
|---|---|
| C7rc | AGCTGGTAACTTCCATGACCAGACCCTCTGCCAAGTGCTGCACATCCATA |
| C8 | AGCTAAAGAAGAGGGGAGGAAACAAGACTAATCAGGAAAGATGAAGGTCTAG |
| C8rc | AGCTCTAGACCTTCATCTTTCCTGATTAGTCTTGTTTCCTCCCCTCTTCTTT |
| C9 | AGCTAAATGGCTTTCACCACTTCCCAGCATCTATTGACATTGCACTCTCAAATATTTTATAAGA |
| C9rc | AGCTTCTTATAAAATATTTGAGAGTGCAATGTCAATAGATGCTGGGAAGTGGTGAAAGCCATTT |
| C10.1 | AGCTTCTATATTCAAGGTAATGTTTGAACCCTGCTGAGCCAGTGGCATGGGTCTCTGA |
| C10.1rc | AGCTTCAGAGACCCATGCCACTGGCTCAGCAGGGTTCAAACATTACCTTGAATATAGA |
| C10.2 | AGCTTGGGTCTCTGAGAGAATCATTAACTTAATTTGACTATCTGGTTTGTGGGTGCGT |
| C10.2rc | AGCTACGCACCCACAAACCAGATAGTCAAATTAAGTTAATGATTCTCTCAGAGACCCA |

| | |
|---|---|
| N1 | AGCTACATTGGCCTGGCTGGTCTCAAACTCCTGACCTTGT |
| N1rc | AGCTACAAGGTCAGGAGTTTGAGACCAGCCAGGCCAATGT |
| N2 | AGCTTCCTGTGTTCAAGCAATTCTAGTGCCTCAGCCTACTTAGTAGCTGGGATGACTG |
| N2rc | AGCTCAGTCATCCCAGCTACTAAGTAGGCTGAGGCACTAGAATTGCTTGAACACAGGA |
| N3 | AGCTCAAAATGCTGGGATTATAGGCATGAGCCACCACCCCCTCCTGGAAGGATTGATATC |
| N3rc | AGCTGATATCAATCCTTCCAGGAGGGGGTGGTGGCTCATGCCTATAATCCCAGCATTTTG |
| N4 | AGCTAATAAATTTTGAAGATAATAAAAGATTTTCACTTATGTTGTCATTTCGGCACAGTTTGGTATAGGATGT |
| N4rc | AGCTACATCCTATACCAAACTGTGCCGAAATGACAACATAAGTGAAAATCTTTTATTATCTTCAAAATTTATT |
| N5 | AGCTGGAAGGCCTCAGCCCCCTCCTCGTACAGCACTGCCTGTTGGAAAGCTTGAGG |
| N5rc | AGCTCCTCAAGCTTTCCAACAGGCAGTGCTGTACGAGGAGGGGGCTGAGGCCTTCC |
| N6 | AGCTATATGAGCTACTTTTATGATTTTATTTTATCCAAAAGAAAGAG |
| N6rc | AGCTCTCTTTCTTTTGGATAAAATAAAATCATAAAAGTAGCTCATAT |
| N7 | AGCTAGGGAAGGAGTAAGGAGACATAAAGGCAATGTGGAGCAGCTGAG |
| N7rc | AGCTCTCAGCTGCTCCACATTGCCTTTATGTCTCCTTACTCCTTCCCT |

3. Forward (f) and reverse (r) primers for the amplification of the apo(a) 1.6 kb intervals.

| | |
|---|---|
| apo(a)16.Sma | TCCCCCGGGATTTACAACTGGGGCTTGG |
| apo(a)16.Nco | TGTTCCATGGTGGGACTGGCCAGCAGT |

4. Primers for generating the deletions of the constructs C1-C10 and N1-N7. Primers ending with `.Sma` and `.Nco` were paired with `apo(a)16.Sma` and `apo(a)16.Nco`, respectively, in the first step of the mismatch-primer protocol. The second-step PCR was carried out with `apo(a)16.Sma` and `apo(a)16.Nco`.

| | |
|---|---|
| apoa.C1.Sma | CCAATGTACTAGGGAGGCTGGAGTATT |
| apoa.C1.Nco | CCCTAGTACATTGGCCTGGCTGGTCT |
| apoa.C2.Sma | GTGCGGCAAGGTCAGGAGTTTGAGAC |
| apoa.C2.Nco | ACCTTGCCGCACTCGACCCTATGTT |
| apoa.C3.Sma | CAGGAGTGGGTGACAGAGCAAGAATG |
| apoa.C3.Nco | CACCCACTCCTGTGTTCAAGCAATTCT |
| apoa.C4.Sma | CATGCCACAGTCATCCCAGCTACTAAG |
| apoa.C4.Nco | TGACTGTGGCATGTTGCCCAGGCTT |
| apoa.C5.Sma | CATTTTGGACCCTGTCTCTACAAAAAATAC |
| apoa.C5.Nco | CAGGGTCCAAAATGCTGGGATTATAGGC |
| apoa.C6.Sma | AAAATTTATTAGATATCAATCCTTCCAGGAG |
| apoa.C6.Nco | TGATATCTAATAAATTTTGAAGATAATAAAAGATT |
| apoa.C7.Sma | CTCATATTTCCCTCAAGCTTTCCAACAG |
| apoa.C7.Nco | GAGGGAAATATGAGCTACTTTTATGATTTTA |
| apoa.C8.Sma | CCCTCACTCTTTCTTTTGGATAAAATAAAA |
| apoa.C8.Nco | GAAAGAGTGAGGGAAGGAGTAAGGAGA |
| apoa.C9.Sma | CTTGAATATCTCAGCTGCTCCACATTGC |
| apoa.C9.Nco | GCTGAGATATTCAAGGTAATGTTTGAACC |
| apoa.C10.Sma | ATGAGAGCTTATAAAATATTTGAGAGTGCAA |
| apoa.C10.Nco | TTTTATAAGCTCTCATGTAAGTCAACAATGT |

```
apoa.N1.Sma       GTGGATCGTGAAACCCCATCTCTACT
apoa.N1.Nco       GTTTCACGATCCACCTGTCTTGGC
apoa.N2.Sma       CACATGCGATGGAGGTTGCAGTGAG
apoa.N2.Nco       CTCCATCGCATGTGCCTCCACACC
apoa.N3.Sma       TATGTTATAAGGAGGCCACGGCAGAT
apoa.N3.Nco       GCCTCCTTATAACATAATTTATAATTACAGAAAA
apoa.N4.Sma       CATCTCCTATTCCTAGTGAACTCACATG
apoa.N4.Nco       TAGGAATAGGAGATGTTAACATTTATACCT
apoa.N5.Sma       AGCCTCCGAGGTTGTGTCGCTAC
apoa.N5.Nco       ACCTCGGAGGCTATGGATGTGCAG
apoa.N6.Sma       TTTCATTCTTGCTGGTAACTTCCATGAC
apoa.N6.Nco       CCAGCAAGAATGAAAGAAGAGGGGAG
apoa.N7.Sma       CCCCCCACCCCTAGACCTTCATC
apoa.N7.Nco       GGGTGGGGGGGAAATGGCTTTCA
```

**Supporting text**

*1. Possibility of sequence composition bias.* We estimated equilibrium base frequencies separately in each region, as the HKY model permits. Although base composition in the exonic versus non-exonic sequence data differed with high statistical significance in all four regions, these differences cannot be explicitly incorporated into an analysis which presumes no foreknowledge of exon location. Furthermore, the use of a ratio-based statistic tends to ameliorate this base-composition bias, which should have roughly the same effect on numerator and denominator. The use of phylogenetic trees allows for properly attributing the contribution to the mutation regimes of more closely-related species subsets (i.e. species belonging to the same family), which could otherwise be overestimated. The tree computation also appropriately weights the number of sequence differences by the evolutionary time available for these differences to occur.

*2. Advantages of using closely-related species for the annotation of genomes.* The fact that all coding sequences share the same gene structure means that a single gene model can be used, thus avoiding the need for probabilistic models that allow for the exon fusion and fission events required in gene finding algorithms of more general scope. The use of closely related species should also facilitate the identification of non-coding regulatory elements. Since these elements are normally short sequences (15-50 bp) containing transcription factor-recognition motifs, increased species divergence makes it more likely that sequence differences, particularly inversions and small insertions and deletions, will complicate their alignment.

## Supporting figures

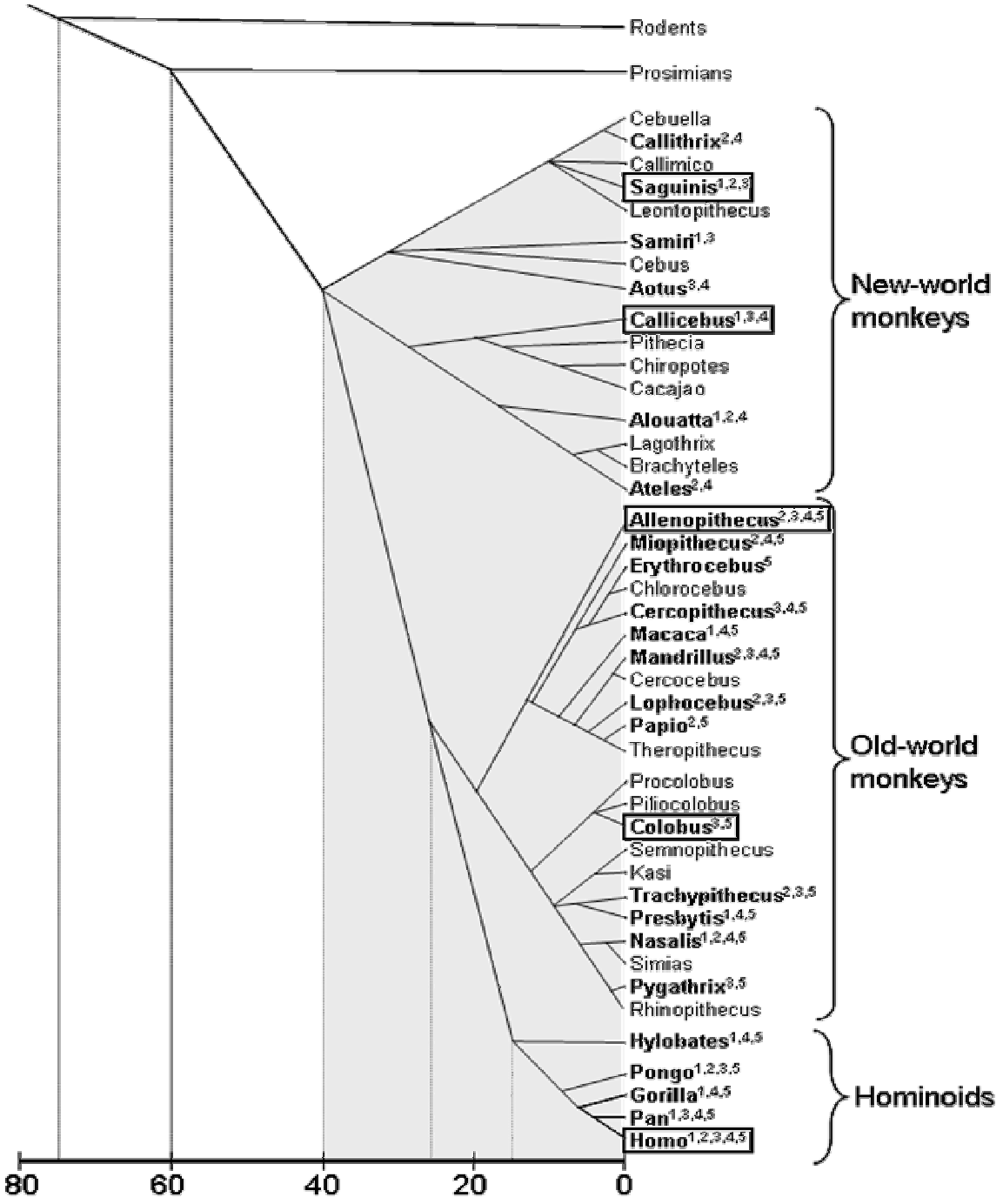


Fig. S1. Primate phylogenetic tree. All the genera composing New- and Old-World monkeys and hominoids are shown. As a reference, prosimians' and rodents' age is also shown. One species from each of the genera in bold was included in this study. Species labeled with 1, 2, 3, 4and 5 were used in the apolipoprotein-B (apo-B), cholesteryl ester transfer protein (CETP), liver-X-receptor-α (LXR-α), plasminogen and apolipoprotein(a) studies, respectively. Species highlighted with a box constituted the optimum subset of five species used to calculate the likelihood curve in Fig. 3. The tree is an adaptation from Fleagle (S5). Age of the taxa is from Goodman (S6).

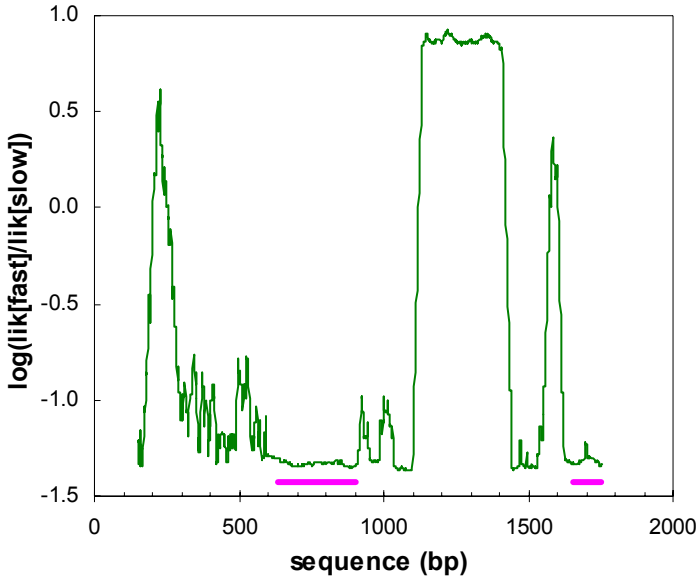


Fig. S2. Likelihood ratios under a fast- versus slow-mutation regime for the genomic interval containing LXR-α exon 3 calculated with the sequence from Homo, Saguinis, Colobus, Callicebus and Allenopithecus (cf. Fig.1). These five species were calculated to contain 72% of the discriminative power of the full multiple sequence analysis of the LXR-α exon 3 region (cf. Table 1). The x-axis represents the position in the multiple alignment consensus sequence, the y-axis the log likelihood ratio at that position. The plot is smoothed using a 20%-trimmed mean over the 50-base window centered at each aligned site. The positions of the complete exon and of a second exon fragment in the sequence are shown by the purple lines under the green ratio curve.

## Supporting references


S1. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res* **22**, 4673 (1994).

S2. G. J. Olsen, H. Matsuda, R. Hagstrom, R. Overbeek, *Comput Appl Biosci* **10**, 41 (1994).

S3. M. Hasegawa, H. Kishino, T. Yano, *J Mol Evol* **22**, 160 (1985).

S4. R. Higuchi, in *PCR protocols* M. A. Innis, D. H. Gelfand, J. J. Sninsky, Eds. (Academic Press, 1990) pp. 177-183.

S5 J. G. Fleagle, *Primate adaptation and evolution* (Academic Press, San Diego, ed. 2nd, 1999).

S6 M. Goodman, *Am J Hum Genet* **64**, 31 (1999).